A Hybrid Clustering and Classification Technique for Forecasting Short-Term Energy Consumption

Mehrnoosh Torabi,[a] Sattar Hashemi,[b] Mahmoud Reza Saybani,[c] Shahaboddin Shamshirband,[d,e] and Amir Mosavi[f,g,h]

[a]Hormozgan Regional Electric Co, Bandarabbas, Iran
[b]Faculty of Computer Engineering, Shiraz University, Shiraz, Iran
[c]Markaz-e Elmi Karbordi Bandar Abbas, University of Applied Science and Technology, Farahani Boulevard, Bandar Abbas, 7919933153, Iran
[d]Department for Management of Science and Technology Development, Ton Duc Thang University, Ho Chi Minh City, Vietnam
[e]Faculty of Information Technology, Ton Duc Thang University, Ho Chi Minh City, Vietnam
[f]Institute of Structural Mechanics, Bauhaus University Weimar, 99423 Weimar, Germany
[g]Institute of Automation, Kando Kalman Faculty of Electrical Engineering, Obuda University, 1431 Budapest, Hungary
[h]Institute of Advanced Studies Koszeg, iASK

This paper presents a hybrid approach to predict the electric energy usage of weather-sensitive loads. The presented method utilizes the clustering paradigm along with ANN and SVM approaches for accurate short-term prediction of electric energy usage, using weather data. Since the methodology being invoked in this research is based on CRISP data mining, data preparation has received a great deal of attention in this research. Once data pre-processing was done, the underlying pattern of electric energy consumption was extracted by the means of machine learning methods to precisely forecast short-term energy consumption. The proposed approach (CBA-ANN-SVM) was applied to real load data and resulting higher accu-racy comparing to the existing models.

Keywords: support vector machine (SVM), artificial neural networks (ANN), electric energy, forecasting, clustering, data mining

INTRODUCTION

Today, the electric power industries play an essential role in the orchestration and progress of fundamental infrastructures of countries. The ever-increasing urban development, popula-tion growth, and industrial expansion, have dramatically expanded the demand for electrical energy with high reliability. Meeting the ever-growing demand for providing energy require massive development of the infrastructures. As there has been no economic method innovated for saving the large amount of electric energy, such a kind of energy should be generated at the same rate as it is being consumed. As a result, it is necessary to identify the potentials of consuming electric energy at the present time and for the future, for the purpose of subtle medium-term and long-term planning for the development of such a kind of industry. One of the ways to satisfy this objective is to aim at empowering the competitive energy markets where electric energy may be distributed between the transaction agencies or broker's markets. Nevertheless, electric energy transaction is always attributed to a certain amount (in MWH) which should be delivered within a certain period. Such a period may be fixed on the basis of the country or the area in which the market is located.

Supply and demand of generation and load should be bal-anced second by second. If such a balance is not achieved, the system might be disrupted and a country or area might encounter power failure [1]. In consideration of the foregoing, there is a need for a managed momentary market which would be able to protect the dependability of the power system and be able to provide techniques for balance of load and genera-tion. Therefore, balance of supply and demand of electric energy is obligatory. In fact, the reliable operation and facilitat-ing in economical dispatch can be considered a goal of power market creation. By taking advantage of various services, which may be conducted by the market, the safety in operation is facilitated. Currently, many countries are already turned into such competitive market for power distribution, while others are considering it. The existing model for this purpose is known as pool, where distributors purchase the electric energy through an accurate prediction strategy [1].

Error of the buyer's consumption need forecast is calculated for every hour of consumption according to Eq. 1.

$$E\ 5\ \eth AC2FC\text{Þ} = AC \quad (1)$$

Where:

AC: Actual Consumption
FC: Forecasted Consumption

Generally, power load forecasting's period can be divided into the following 4 groups [2]:

Very short-term load forecasting (STLF): this includes information on power load for any period from a few minutes to a few hours;

STLF: this includes forecasting every 30 minutes or every hour within a day or a week;

Mid-term load forecasting: this includes weekly or monthly power load forecasting within a year or up to 3 yr;

Long-term load forecasting: this includes weekly, monthly and yearly forecasting within a period of time greater than 3 yr;

One of the short-term load forecast targets is to foresee the electric energy consumption for power market in the upcoming days.

Short-Term Load Forecasting

Short-term load forecasting (STLF) plays an important part in the economy, real-time control of the system and safely commissioning of an electric power system [3,4], because it provides necessary information for distribution of economic load. In Iran's power market, STLF is used. Buyers are required to announce their estimated hourly demands at least 3 days prior to the market's opening [5]. A wide range of advanced data analysis tools have been used from real engineering application to medical sciences since the 1990s
[6]. Researchers, ever since, have been keen in the advance-ments of methods and tools for getting insight and identify-ing patterns in complex data to predict future conditions [6].

Data mining is a popular theoretical notion that effectively aids industrial practitioners in coming up with viable solu-tions for forecasting problems. Data mining may benefit from the advanced (machine learning) ML tools, where the histori-cal data are used to train and create the perdition model. Such learning can be identified through finding meaningful patterns and correlations between variables. In this context, data science technologies such as data mining aim at getting insight and identify hidden patterns and correlations between variables to predict the forthcoming situations [7]. In the context of energy market, the accurate load prediction is extremely crucial in designing new power systems. As mentioned above, the buyers are required to report their predicted demands at least 3 days ahead of the market's opening. If the percentage of forecast error is greater or less than the permissible amount set by the market, heavy fines will be considered for the buyer, thereby making it necessary to predict the electric energy consumption. Because the elec-tric energy consumption trend has predictable daily and weekly changes [1], it can be predicted by data mining.

A number of modeling techniques such as auto and linear regression moving average, as well as time series have been introduced in the last few decades, a for the STLF [8–10]. Generally, these methods can be studied in two different models categories of dynamic and static. Within a static model, the load is seen as merely explicit time functions of polynomial or sinusoids where the effect of weather are not considered in planning [11]. Since there is a strong correla-tion between the behavior of power consumption and weather variables, common problems with these approaches include the inaccuracy of prediction. Yet, the dynamic mod-els employ significant variables and effective parameters for achieving higher level of robustness and accuracy. Neverthe-less, there exists a number of certain prediction models which do not involve weather data such as researches pre-sented in e.g., [12–14], while other prediction models include the effects of weather variables. In this context, among dif-ferent forecasting methods in recent years the artificial neural networks (ANNs) have received particular attention [15–17]. This popularity has been due to its capability of supervised learning of non-linear functions. Note mentioning that ANN methods were used in different engineering applications and their popularity is not limited to load forecasting [18–21].

Paras Mandal et al. (2006) [22] used weighted Euclidean method to determine similarity between forecasting day and previous days, and then they used feed-forward and back propagation neural network to forecast the load. Philippe Lauret et al. (2008) [23] used Bayesian neural network techni-ques to forecast short time load. In this technique different models are compared and the optimal model with the most important input variables is selected; this leads to automatic tuning of regularization coefficients of the model. In another research, Dan Jigoria-Oprea et al. (2009) [2] used recursive ANN for STLF.

Further machine learning modeling techniques for robust modeling includes support vector regression and support vector machine (SVM) which have shown promising perfor-mance in many applications [10,21], including predicting the load [24]. Torabi et al. [25] applied SVM and ANN to predict an hourly electric energy consumption in Iran. Pereira et al. (2006) developed a model based on SVM for short-term load prediction where the average daily loads are clustered utiliz-ing threshold values in terms of patterns [26]. Later, Jain et al.
[12] used "load at the same time" and loads for the "same hour of the previous 1–7 days" as inputs, and advanced an hourly load prediction model using a SVM algorithm. PSO3-based SVM was used by Sun et al. (2006) [27]. PSO is used to automatically select input variables for the SVM model. In another research, Dongxiao Niu et al. (2010) [28] used Ant Colony Optimization-based SVM (ACO-SVM). In their model, ACO was used to preprocess datasets including outlier elimi-nation and feature selection. In addition, Nie et al. (2012)
[29] used a hybrid model of SVMs and ARIMA for short-term load estimation.

This research aims at predicting the energy demand of the short-term for the Bandarabbas region in Iran. The research focuses mainly on building models based on the preprocessed data. In contrast, CRISP methodology, that is one of data mining processes, indicates that the preprocess-ing of data is one of the most important and effective phases of data mining. This is different from that of earlier ones in the sense that it had special focus on both building the model and preprocessing of data. Through this research, a new idea with which the accuracy of load forecasting can be improved is presented. A novel cluster-based approach using ANN and Support Vector Machine (CBA-ANN-SVM) is created for energy consumption forecasting. This study maintains positive contributions of previous studies [25], and in the meanwhile incorporates the following contributions into forecasting process:

1. Studying the impact of special weather condition of Bandarabbas and its impact on hourly, weekly and seasonal consumption.
2. Presenting other effective parameters and studying their impact on model in details.
3. Including the parameter "power outage" in the calcula-tion of total energy.
4. Indicating forecasting accuracy is higher than ANN and single SVM.

The 18 months of data in the years of 2010 and 2011 is considered in this study. In Section 2 we show background of the CRISP process and the algorithms that were used. Effective data on energy consumption can be understood and preprocessed in section 3. Furthermore, we consider the diagrams for visualizing patterns of the electrical energy con-sumption. Proposed forecasting methodology is presented in section 4. We describe three models of SVM, ANN, and CBA-ANN-SVM (a novel hybrid model of clustering with both widely-used SVM and ANN) to predict the electric energy. Finally, we will see experiments in section 5.

BACKGROUND

In this section, we first introduce the preliminary concepts and algorithms which are used.

Data Mining Process

There are different ways of implementing data mining tasks; one powerful method is Cross Industry Standard Process (CRISP) for Data Mining. CRISP process consists of 6 phases intended as a cyclic process [30]: business understanding, data understanding, data preparation, modeling, evaluation, and deployment.

Description of Used Algorithms

Support Vector Machine

SVM is a distinguished supervised classifier and based on statistical learning theory [31]. This method maps the original input space through multiple dimensional space [32]. This method minimizes learning error by minimizing structural risk mapping the original finite-dimensional space into a much higher-dimensional space where linear learning algorithms can be used [33]. When SVM is faced with sample data in high-dimensional, it performs well. SVM has a significant number of applications and methods in the field of load forecasting because it is highly generalizable and a sparse solution repre-sentation; it also has no problems with local minima [4].

Artificial Neural Networks

One of the strong points of ANN is its ability to learn non-linear and complex relationships [17] that are difficult to model using conventional methods. It is used for solving different types of problems by simulation with limited gathered data. If a neural network is trained well, its speed in modeling is much faster than any other numerical methods [34]. Further advantages of ANN is reported to be the tolerance against invalid input dataset, and also the capability in finding patterns under unsupervised learn-ing conditions. When there is little information about features, ANN is a suitable choice. ANN is used frequently in STLF with good/satisfactory performance. It can model an unspecified non-linear relationship between load and weather variables [35].

K-Means

One of the well-known clustering algorithms that is widely studied is K-means [36]. It is a simple method to parti-tion data into specified number of clusters (k). The applica-tion of K-means is widespread, because it is faster than hierarchical clustering, its implementation is simple and it is easy to use [37]. This algorithm performs well when clusters are compact clouds that are well separated from each other.

One of the data mining problems that is widely studied is clustering using the k-means

DATA UNDERSTANDING AND DATA PREPARATION

Data Understanding

One of important issues of short-term energy forecasting is selecting input variable. The goal is to find a relationship in terms of distance and time among input variables to find a suitable rela-tionship between input and output variables. There are no gen-eral rules for selecting input variables. This selection depends much on one's experience, experts' views, and preliminary tests; although some statistical data analysis can be useful for determin-ing which variables have significant impact on consumed energy. This research uses two different types of data [35]:

Energy

Hourly consumption data

Supplied from substations across the research area using electronic counters. It includes the total hourly values as a single value for every hour.

Zero consumption data

The possibilities of power outage across the research area. Zero consumption due to the power outage, effects on amount of forecasting, because it would not be the consid-ered value of consumption in the energy consumption record. So, to calculate the amounts of electrical energy cor-rectly, not only the real amounts of energy consumed should have been considered, but also the amounts of energy due to power outage considered.

Amount of power outage is calculated according to Eq. 2:

$$W = PT \quad (2)$$

$$T = \frac{T_1 2 T_2}{60} \quad (3)$$

Where:

W: Not used electric energy (KWH)

P: Power (KW)

T: Time (H)

T_1: On time

T_2: Off time

T_1-T_2 is minute

Because the cycle of generation, transmission and distri-bution of electrical power are flowing through balanced 3 phase circuits, the amounts of active power in balanced 3 phases can be calculated from Eq. 4.

$$P = \sqrt{3} V_L I_L \cos\varphi \quad (4)$$

Where:

V_L: Effective line voltage (kW),

I_L: Effective line current (A)

u: Phase angle between phase voltage and phase current, it is also known as impedance angle

Cos u: power factor, it is equal to 0.9 in Hormozgan transfer network.

Weather Data

Weather data include hourly measured air temperature, humidity, wind direction, wind speed and atmospheric con-ditions. This data is collected from the weather website "wunderground" [38].

Perceived temperature

Through analyzing the dataset, another important parame-ter has been discovered as "perceived temperature". In fact, the warmth that the average person perceives in different time of the year is somewhat felt warmer or colder than the actual temperature measured. Thus, the perceived tempera-ture is the warmth or cold a typical healthy human may feel. The index is presented as an integration of "wind chill factor" and "heat index" referred also as outdoor tempera-ture. The wind chill factor is the temperature that human feels when exposed to wind as a function of wind speed and temperature. In the cases when wind speed exceeds 5 km/h and temperature is below 48C and the, the wind chill factor can be obtained as follows [39]:

$$W = 13.1210.62153T - 11.373V^{0.16} + 10.39653T \cdot V^{0.16} \quad (5)$$

Where:

W: Wind chill factor (8C)

T: Actual weather temperature (8C)

V: Wind speed (km/h)

To obtain the temperature that human feels the heat index is utilized when the weather temperature is above the body temperature. And, in fact, that is when the body starts to sweat to cool down from heat. Conversely, when the rela-tive humidity increases, the rate of evaporation decreased, resulting in the body store more heat. As the result, the body feels warmer when the humidity is higher. Consequently, the heat index can be considered as an integrated function of

Table 1. Calculated Pearson correlation coefficient between consumption and variables to prove the effectiveness of these parameters in electrical energy consumption.

	Wind	Feels	Last day Average					Last Hour	2 Last Hour	3 Last	Last day Usage	Last Week
	Temperature		Mumidity	Speed	Like	Temperature		Usage	Usage	Hour Usage	Same Hour	Usage Same Hour
Pearson Correlation	0.		0.	0.	0.	0.8223148694		0.	0.	0.999896	0.	0.9999

relative humidity and the actual temperature. For the air temperature above 278C and the humidity more than 40% [40], the heat index is obtained according to the Eq. 6 [41] [25]:

$$HI\ 5c_1 1c_2 T\ 1c_3 TR 1c_5 T^{\ 2} 1c_6 R^2 1c_7 T^{\ 2} R 1c_8 TR^2 1c_9 T^{\ 2} R^2 \qquad (6)$$

$c_1 5242{:}38$; $c_2 52{:}049$; $c_3 510{:}14$; $c_4 520{:}2248$;

$c_5 526{:}383310^{23}$; $c_6 525{:}482310^{22}$;

$c_7 51{:}228310^{23}$; $c_8 58{:}528310^{24}$; $c_9 521{:}99310^{26}$

Where:

HI: Heat index (8F)

T: Actual temperature (8F)

R: Relative humidity (%)

Note that when it is not possible to estimate the heat index and wind chill factor, they will be considered to be equal to the actual temperature.

Last day temperature

Since air temperature is the most important atmospheric parameter that has impact on the amount of power con-sumption, and also air temperature of the past affect the load curve, minimum, maximum, and average air temperature are calculated. Correlation coefficients of these three values on consumed energy are presented in Table 1. As illustrated, correlation coefficient of average temperature is higher than that of minimum and maximum temperatures, therefore only average air temperature of the previous day is used in building the model in this study.

Data Preparation

One issue of data preparation is data cleansing. This task includes: filling in missing or incomplete values with appro-priate values, identifying outliers and removing them, remov-ing duplicates, and taking care of incorrect data types. Poor data preparation results in incorrect and unreliable data min-ing results, and the discovered knowledge will be of poor quality.

In Power Company, the reports regarding consumption of electric energy and blackout may be prepared on a weekly basis. At the time of preparation of such reports, a large amount of errors (outliers, faulty values and missing values) may be resolved.

One of the useful methods for visual inspection of data is Line Chart [42]. Such a method is useful for finding the noisy data in a one to three-dimension space [43]. Therefore, the hourly diagrams of electric energy consumption were drawn on a weekly basis. Such an action was taken for hourly dia-grams of temperature, humidity, and wind speed. Noisy data was completely distinguished in the diagrams. There was a low amount of noisy data in comparison with the entire data. After consultation with the experts of Power Company (for the data of electric energy consumption) and the experts of Weather Bureau (weather data) it became clear that the noisy data are the faulty values data. Such data was corrected as the method of correcting missing data, which is explained as follows. For example, in the Figure 1, the diagrams of consumption of electric energy in the 1st week of February, April, May, June, July, August, October, and December for the year 2011 are presented (X axis representing time on the basis of Day and Y axis representing energy consumption on the basis of MWH).

In the drawn diagrams, the faulty values data is distin-guished, because it takes out the figure of the diagram from the expected course. Method of behavior with the faulty val-ues is alike the behavior with the missing data.

One of the methods for filling the missing data is the use of attribute mean [42,44]. In consideration of Figure 1, the rate of changes of energy consumption from one hour to the

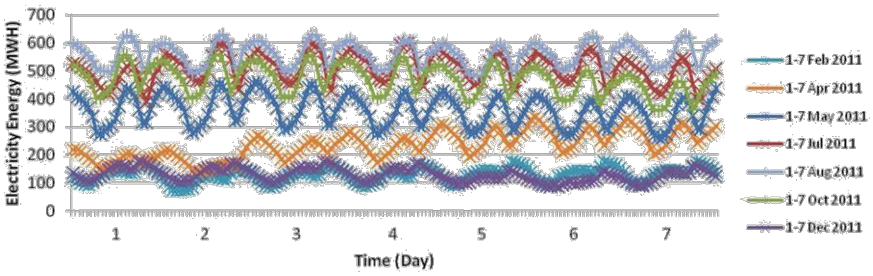

Figure 1. Hourly diagrams of electric energy consumption in the 1st week of February, April, May, June, July, August, Octo-ber, and December for the year 2011 (X axis representing time on the basis of Day and Y axis representing energy consump-tion on the basis of MWH). [Color figure can be viewed at wileyonlinelibrary.com]

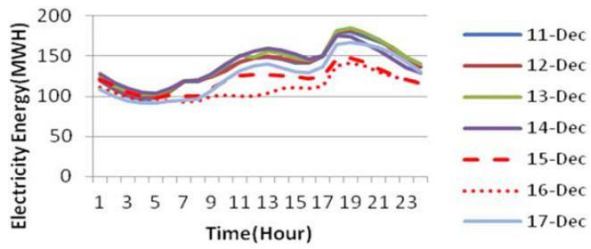

Figure 2. Electricity consumption 11–17 December 2010 (15th and 16th are holiday). [Color figure can be viewed at wileyonlinelibrary.com]

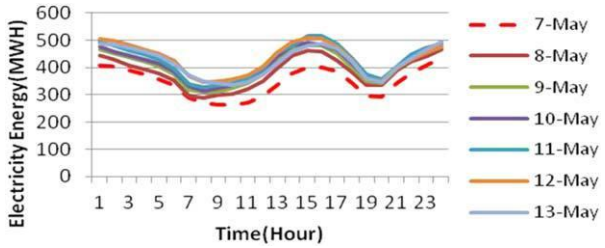

Figure 3. Electricity consumption 7–13 May 2011 (7th day is holiday). [Color figure can be viewed at wileyonlinelibrary. com]

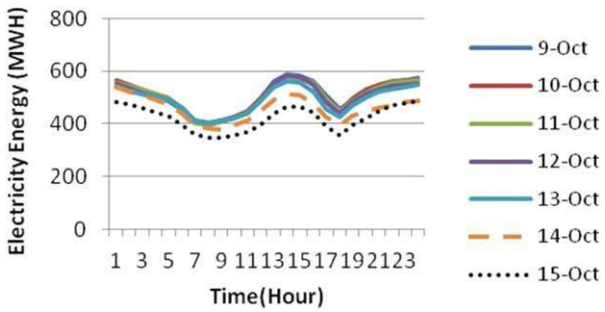

Figure 4. Electricity consumption 9–15 October 2010 (14th and 15th are Thursday and Friday(weekend)). [Color figure can be viewed at wileyonlinelibrary.com]

next is a low rate. The changes in the diagram gradient in every point in comparison with the previous point (consumption in the previous hour) and the next point (consumption in the

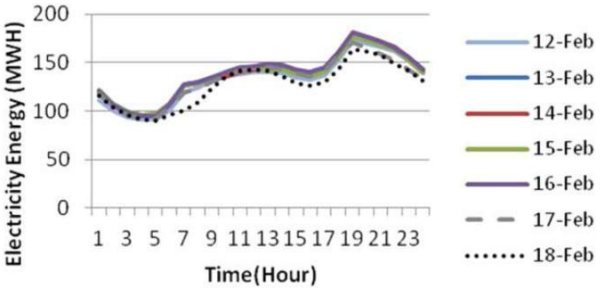

Figure 5. Electricity consumption 12–18 February 2011 (17th and 18th are Thursday and Friday (weekend)). [Color figure can be viewed at wileyonlinelibrary.com]

next hour) is a low rate and there is no sudden change. In the diagrams 2 to 5 (such 4 diagrams are explained by detail in sub-section 3-3), as the time interval is one day instead of one week, it may become specified that the rate of changes is low. Because of the nature of the data, for filling the missing values, we may replace them with average of consumption in the pre-vious hour and next hour, and as the number of the missing values is not noticeable, such a method is very useful. For such a purpose, we put to order the energy consumption data on the basis of time, and then we filled in the missing values under the method mentioned. As the climate conditions of Bandarab-bas is relatively stable, there is no sudden change, and as the number of missing data is low, therefore, for filling in the miss-ing data regarding wind direction and condition, we replaced such missing data with the data from the previous hour.

To be able to use various data from different data sources such as weather information, power outage, and the amount of power consumption, they needed to be integrated. For this purpose, MS-SQL Server 2008 was used to store the data, the integration was done by using unique ID for common variables and by converting date from Iranian Calendar to Gregorian Calendar or vice versa.

Load Pattern Recognitions

Research reveals that there usually exists specific weekly pattern for every hourly load pattern [45]. To find a trend and governing rules for electricity consumption, Figures for daily electricity consumption for all seven days of the week are illustrated. (In Iran, week days start from Saturday until Wednesday; Thursdays and Fridays are considered the weekend).

In the following, some Figures are presented; dashed and dotted lines correspond to holidays. In each Figures 2–5 4 diagrams are drawn for 7 weekdays (Saturday to Friday). Each diagram shows the electric energy consumption within

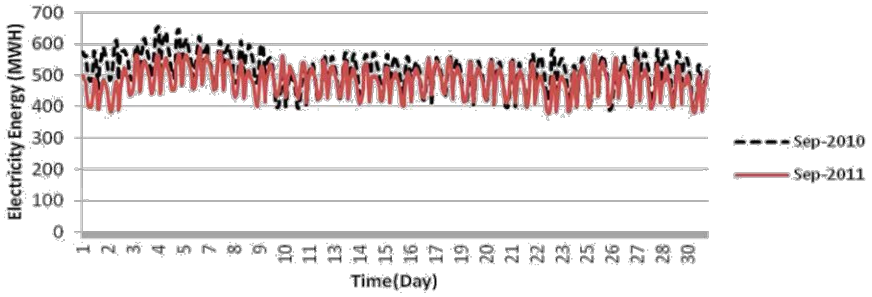

Figure 6. Electricity consumption for September of 2010 and 2011. [Color figure can be viewed at wileyonlinelibrary.com]

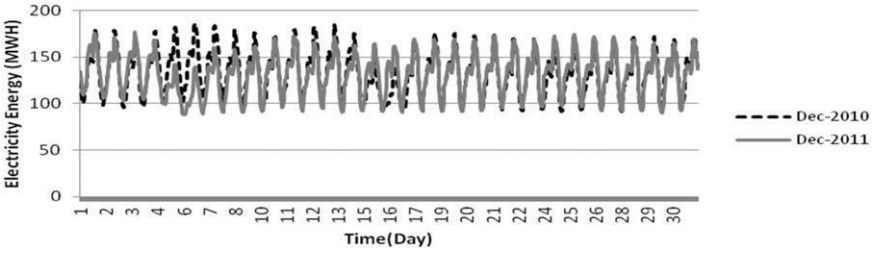

Figure 7. Electricity consumption for December of 2010 and 2011.

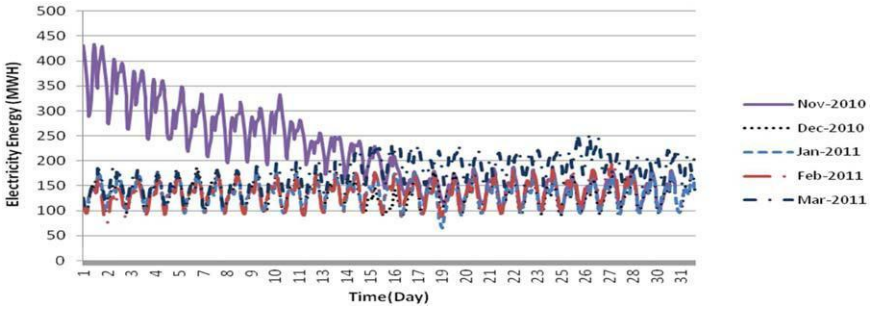

Figure 8. Consumption pattern for cold months of the year; Cold months include 11th, 12th, 1st, 2nd, and 3rd months of the year. [Color figure can be viewed at wileyonlinelibrary.com]

24 h of night and day. Ordinate (X axis) shows time by hour and abscissa (Y axis) shows the rate of electric energy con-sumption by MWH. To specify the trend of consumption in various days, the diagram of every weekday is drawn with a different color. In Figures 2 and 3, the diagram of the holi-days is shown with red color. If two days are closed, one day would be shown in the form of broken line and the other day is shown in the form of a dotted line. In Figure 2, December 15 and December 16 are holidays. In Figure 3, May 7 is holiday. By paying attention to such 2 Figures, we may see that the rate of consumption on holidays is less than other weekdays, because, the width of an almost unani-mous majority of the dots in the red diagrams (holidays) is less than the width of the dots in other diagrams.

As can be seen from the Figures, power consumption during week days are very different from weekends. Depending on the cold or warm periods, this impact can either affect only consumed or also affect trend and con-sumption pattern. In Figure 4, October 14 and October 15 are weekends. In Figure 5, February 17 and February 18 are weekends. In such 2 Figures, the diagrams of the weekends are shown in the form of a broken line for Thursdays and, shown in the form of a dotted shape for Fridays. Figure 4 pertains to the warm month (October). We may see that the weekend is effective on the consumption rate, because the consumption is decreased toward the workdays. Whereas, in the Figure 5, which concerns the cold month (February), in the weekends, not only the consumption rate is decreased, but also the consumption pattern is changed in some hours.

Analysis of consumption trend for the same months of different years have shown that consumption pattern in all cases are nearly similar. To show such a matter, the diagram of energy consumption in the months September, and December of the years 2010 and 2011 is drawn (Figures 6 and 7). For instance, in Figure 6, the diagram of electric energy consumption in September (from the 1st day to the 30th day), is drawn in recognition of the years 2010 and 2011. X axis represents the time on the basis of Day, and Y axis represents consumption of electric energy on the basis of MWH. According to such Figures, the consumption way of the same months in two different years are approximately similar. In such cases, not only the consumption trends but also the consumption rates are the same.

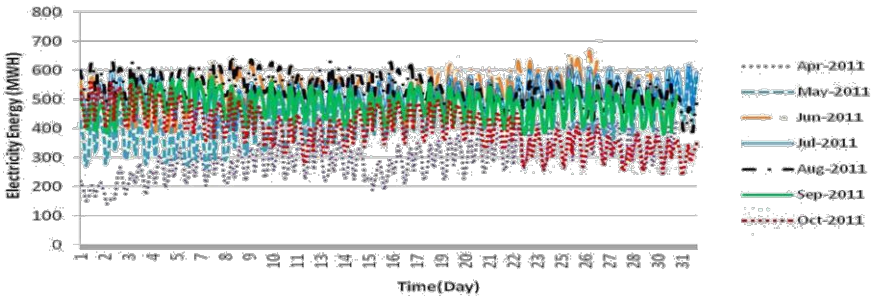

Figure 9. Consumption pattern for warm months of the year; Warm months are from the 4th month to the 10th month of the year. [Color figure can be viewed at wileyonlinelibrary.com]

In Figures 8 and 9, the diagram of electric energy con-sumption in 12 months of a year is drawn on the basis of the months. In Figure 8, the diagram of the cold months is drawn set close together and in Figure 9, the diagram of the warm months is drawn close together. X axis represents the time on Day basis (from the 1st day of a month to the last day of the month), and the Y axis represents the rate of elec-tric energy consumption on MWH basis. As illustrated in Figures 8 and 9, it is obvious that power consumption has some seasonal patterns. Based on the consumption trend, months of the years can be grouped into cold and warm months. Cold months include 11th, 12th, 1st, 2nd, and 3rd months of the year, and warm months are from the 4th to the 10th month of the year.

Correlation of Features and Their Importance on Electric Energy Consumption

As mentioned earlier, based on opinion of experts and previous studies the essential parameters in electric energy consumption are considered humidity, temperature, and weather condition [22,46]. Accordingly, a number of effective parameters are emphasized, i.e., the hourly energy usage and three hours, and the relevant historical data of the last day, week, and month whether week days or weekends. To demonstrate the efficiency of these parameters in electrical energy consumption, correlation coefficient (CC) of these parameters and output variable (electrical energy consump-tion) have been calculated and checked by using SPSS soft-ware (PASW Statistics 18 Release 18.0.0). Therefore, the CC is adopted to measure the correlation of two numeric parame-ters. Here the value of CC is between 21 and 1. The value of CC between two random variables x and y is calculated according to Eq. 7.

$$q_{x,y} = \frac{Covðx; yÞ}{r_x r_y} \quad (7)$$

Where:

Cov(x,y): Covariance of variables x and y

r_X : Standard deviation of x

r_Y : Standard deviation of y

Table 1 presents the correlation coefficient between con-sumption and variables.

As shown in Table 1, correlation coefficient for tempera-ture, feels like, minimum, maximum and average tempera-tures of previous day are high. Since correlation coefficient between average temperature of the previous day and mini-mum and maximum temperatures are very high, one can ignore minimum and maximum temperatures of the previous day and consider only their average values. Correlation coef-ficient for humidity, wind speed is very low, thus their influ-ence in building the model is very low. Therefore, their impact on the model is very minimal. In fact, their effects are indirectly used in building the model. Humidity and wind speed effect feel-like temperature.

PROPOSED METHOD

Here, the Clementine v. 12.0 is utilized for building the model. Three different methods were used to build the model for this research: SVM, ANN, and CBA-ANN-SVM. In the following, all models are explained in details, and the best model with the least forecasting error is selected.

SVM Approach

Here, SVM as an effective machine learning is used per-forming well when few instances are available, classes are not linearly separable, dataset is high-dimensional or there exist local minima. If there is a small set of support vectors, then SVM is capable of high generalization.

For modeling, initially we have to bring data into Clemen-tine. Source node that has been named “Imported Data”, reads in data from external source (dataset that we have prepro-cessed) into Clementine. We used a “Partition” node to split the data into separate subsets or samples for training and evalua-tion stages of model building. This node has been set to use 90% of the data for the training and the remaining 10% for the testing. Partition node has “random seed” option. With this option, we can ensure a different sample (by selecting another subset of data records) will be generated each time the node is executed. By “Type” node, we tell Modeling node (“SVM” node) whether fields will be predictor fields or predicted fields. This node also describes data type (string, integer, real, date, time, or timestamp) in a given field. “SVM” node is a Modeling node. This sequence of operations is known as a data stream. When the stream is executed and model is built, the model nugget is created and added to the Models palette in the upper right corner of the application window. In accordance with Clementine software, to see modeling result we have to add the model nugget to the stream and attach the model nugget to the “Type” node, at the same point as the Modeling node. “Analysis” node helps to determine whether the model is acceptably accurate.

Building the SVM model requires a trade-off between maximizing the margins and the learning error. The used software (Clementine) has a regularization parameter “c”, which is used to regulate this trade-off. Increasing c leads to higher classification accuracy (reduced regression error) however it may also lead to overfitting. This study tested fol-lowing three kernel functions: linear, sigmoid and polyno-mial. Various combinations of inputs were used to build the model. During each model building, the importance of input parameters was verified by calculating Mean Absolute Per-centage Error (MAPE) and variance. Some inputs had very little impact on the MAPE, and therefore they were omitted from the list of input variables. MAPE is obtained from the Eq. 8 [10].

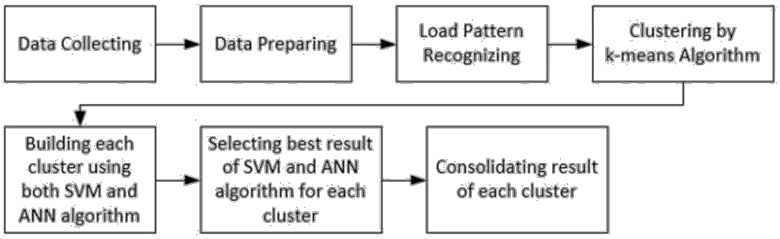

Figure 10. flow chart of CBA-ANN-SVM model.

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\frac{|P_A^i - P_F^i|}{P_A^i} \times 100 \quad (8)$$

Where:

P_A^i: Real load

P_F^i : Predicted load values

N: Number of instances

List of input data that was used in the model:

1. Consumed energy of previous hour
2. Consumed energy of previous two hours
3. Consumed energy of previous three hours
4. Consumed energy at same hour of previous day
5. Consumed energy at same day and same hour of previ-ous week
6. Month of consumption
7. Air temperature of the same hour
8. Feels-like temperature
9. Average air temperature of previous day
10. Hour of consumption
11. Weather condition
12. Weekend

ANN Approach

The second approach for our model is the use of ANN. All stages are like SVM approach except Modeling node that "ANN node" is used.

CBA-ANN-SVM Approach

Another model that was used in this study was based on clustering; the goal is to verify the strength of clustering for forecasting. CBA-ANN-SVM consisted of two phases. In the first phase, parameters such as amount of energy consump-tion, month, weekend, and holidays were used in an unsu-pervised learning, inputs were clustered into several subsets that were similar to each other. In the second phase, each cluster (subset) was assigned to a supervised SVM and ANN to forecast power consumption. K-means was used to cluster the dataset. The application of K-means requires the user to give the number of desired clusters as input parameter to the algorithm.

According to subsection 3-3, in Figures (2 and 3), con-sumption patterns for weekends and holidays are different from week days, therefore maybe we have one cluster for weekends and holidays and another cluster for weekdays

In Bandarabbas, Thursdays and Fridays are the weekend. On Thursdays, only the schools and most of the departments are closed, but on Fridays and holidays, all schools, all departments, private companies, factories, plants, banks, business centers, etc. are closed. So, there is a possibility to classify the Thursdays in one cluster, and the Fridays and holidays in another cluster. In consideration of the Figures 8 and 9, the consumption patterns in the cold months (the 1st to the 3rd month and the 11th to the 12th month) are similar to each other and the consumption patterns in the warm months (the 4th month to the 10th month) are similar to each other, and such two patterns are distinguishable. Then, probably, the cold months are in one cluster and the warm months are in another cluster.

As a result, there is a possibility to be a cluster for the workdays in warm months and another cluster for the work-days in the cold months. It is possible that all holidays and weekends to be in one cluster, or be in two different clusters according to the cold and warm months. Therefore, we may have 3 or 4 clusters. If the cluster of the Thursdays and the cluster of the Fridays and holidays are different from each other, then we have 6 clusters. In the recent clustering, the clusters are as follows:

Workdays of warm months
Workdays of cold months
Thursdays of warm months
Thursdays of cold months
Fridays and holidays of warm months
Fridays and holidays of cold months

In k-means algorithm, we regulated the number of clus-ters to 3, 4 and 6 and analyzed and examined the rules related to the clusters. When the number of clusters was 6, the rules had no special meaning, but, when the number of clusters was 3 and 4, the rules were completely meaningful:

Three clusters:

Cluster 1: Weekend, and other holidays.

Cluster 2: Saturday through Wednesday of warm months (work days).

Cluster 3: Saturday through Wednesday of cold months (work days).

Four clusters:

Cluster 1: Saturday through Wednesday of cold months (work days).

Cluster 2: Holidays.

Cluster 3: Thursdays.

Cluster 4: Saturday through Wednesday of warm months (work days).

Based on the identified patterns in subsection 3-3, this algorithm has performed well in both clustering cases.

Having done the clustering, models of each cluster were built separately using SVM or ANN. Before modeling, data-sets had been divided into learning (90% of data) and testing data (10% of data) by using partition node, as we had in the two previous methods. To obtain the final error of models, results of each cluster were consolidated by append node. This stream was executed 10 times.

After clustering by K-means algorithm that sets to 4 (3) clusters, each cluster divided into learning (90%) and testing (10%) data by using Partition node. Models of each cluster were built separately using both SVM and ANN. Each of SVM or ANN models that had better results (less MAPE and SD), were selected as the final model of each cluster. Finally, results of each cluster were consolidated by Append node.

Inputs in section 4-1, except weekends (because week-end parameters had been used previously for clustering), were used for both models.

Figure 10 shows the flow chart of CBA-ANN-SVM model.

Table 2. MAPE and SD of CBA-ANN-SVM (4 Clusters) model.

	Model	Description	MAPE	SD
Cluster1	SVM	d 5 2, c 5 15, g 5 4	1.279	1.800
Cluster2	ANN	IL 5 1, HL 5 2, OL 5 1	1.621	1.919
Cluster3	SVM	d 5 2, c 5 15, g 5 2	2.112	1.927
Cluster4	SVM	d 5 1, c 5 20, g 5 5	1.452	2.001
Total	-	-	1.476	1.722

d: Degree of Polynomial.
c: Regularization parameter.
g: Gamma.
IL: Input Layer.
HL: Hidden Layer.
OL: Output Layer.

Table 3. MAPE and SD of CBA-ANN-SVM (3 Clusters) model.

	Model	Description	MAPE	SD
Cluster1	SVM	d 5 4, c 520, g 5 2	1.199	2.214
Cluster2	ANN	IL 5 1, HL5 1, OL 5 1	1.539	1.005
Cluster3	SVM	d 5 2, c 515, g 5 24	1.071	1.583
Total	-	-	1.297	1.690

d: Degree of Polynomial.
c: Regularization parameter.
g: Gamma.
IL: Input Layer.
HL: Hidden Layer.
OL: Output Layer.

EXPERIMENTS

SVM Approach

According to the 10 times 10-fold cross-validation method [48], the data stream is executed 10 times. By using random seed option, a different sample will be generated each time the stream is executed. "Analysis" node verifies whether there is an over-fitting or not and checks the value of MAPE and Standard Deviation (SD). The final model was built using a grade 3 polynomial function and c, were set to 20 and 4 respectively.

Average result of MAPE and SD of 10 times 10-fold cross-validation is:
MAPE: 2.015
SD: 2.101

ANN Approach

The data stream of this model is similar to SVM model. In a supervised ANN, each phase of learning is called a cycle. Cycles continue until networks' weight become stable. The parameter "Persistence" is set to 1000 in this model, meaning that if in 1000 cycles the error remains constant, then the model has become stable. In this approach, 10 times 10-fold cross-validation method is used, too. The best model was built with one layer of input, two hidden layers, one output layer and input parameters in section 4-1. Average result of MAPE and SD of 10 times 10-fold cross-validation is:
MAPE: 1.790
SD: 1.971

Table 4. Results for comparison of proposed Models, SVM, ANN, and CBA-ANN-SVM (3 and 4 clusters).

			CBA-ANN-SVM	
	SVM	ANN	4 Clusters	3 Clusters
MAPE	2.015	1.790	1.476	1.297
SD	2.101	1.971	1.722	1.690

CBA-ANN-SVM Approach

Tables 2 and 3 show used models for this study alongside the average results of MAPE and SD of 10 times 10-fold cross-validation for each cluster. It is important to notice that the impact of MAPE and SD of each cluster is not the same; this is due to variation in numbers of inputs. As can be seen in the tables, when dataset is clustered in 3 clusters, the fore-cast is better. When we have 4 clusters, holidays and week days are in different clusters. But with 3 clusters, because the patterns of these groups are very similar, they will be placed in one cluster. In the result of merging these two clusters, training data will be increased. The more the training data with similar properties, the better the model.

Comparison of Models

Table 4 shows the results of all three models. Least average percentage error belongs to CBA-ANN-SVM model with three clusters. Therefore, this model was selected as the final model.

Comparison of Results Obtained with the Proposed Model and with Similar Published Methods

The MAPE of the proposed model, the CBA-ANN-SVM, with three clusters, is 1.297. The MAPE of Hooshmand's model [17] is 1.603. This paper used a two-step algorithm. At first a wave-let transform and ANN were used for the primary load forecast-ing and then the "similar-hour" method and adaptive neural fuzzy inference system were used to improve the results of the primary load forecasting. Nie et al. used a hybrid model of ARIMA and SVM. Initially they used ARIMA to forecast the daily load and then they used SVM to correct the deviation of any former forecasting. MAPE of this model was 3.85.

CONCLUSION AND FUTURE WORK

Electrical energy distributor companies in Iran have to announce their energy demand at least three 3-day ahead of the market opening. Therefore, an accurate load estimation is highly crucial. This research invoked methodology based on CRISP data mining and used SVM, ANN, and CBA-ANN-SVM (a novel hybrid model of clustering with both widely used ANN and SVM) to predict short-term electrical energy demand of Bandarabbas. In previous studies, researchers introduced few effective parameters with no reasonable error about Bandarabbas power consumption. In this research we tried to recognize all efficient parameters and with the use of CBA-ANN-SVM model, the rate of error has been minimized.

After consulting with experts in the field of power con-sumption and plotting daily power consumption for each week, this research showed that official holidays and week-ends have impact on the power consumption. When the weather gets warmer, the consumption of electrical energy increases due to turning on electrical air conditioner. Also, con-sumption patterns in warm and cold months are different. Ana-lyzing power consumption of the same month for different years had shown high similarity in power consumption pat-terns. Factors with high impact on power consumption were identified and statistical methods were utilized to prove their impacts. Using SVM, ANN and CBA-ANN-SVM, the model was built. Sine the proposed method (CBA-ANN-SVM) has low MAPE 5 1.474 (4 clusters) and MAPE 5 1.297 (3 clusters) in

comparison with SVM (MAPE 5 2.015) and ANN (MAPE 5 1.790), this model was selected as the final model. The final model has the benefits from both models and the benefits of clustering. Clustering algorithm with discovering data struc-ture, divides data into several clusters based on similarities and differences between them. Because data inside each cluster are more similar than entire data, modeling in each cluster will pre-sent better results.

For future research, we suggest using fuzzy methods and genetic algorithm or a hybrid of both to forecast each cluster. It is also possible to use fuzzy methods or genetic algorithms or a hybrid of both without using clustering. It is issued that such models will produce better and more accurate results.

LITERATURE CITED

1. Kirschen, D., & Strbac, G. (2004). Fundamentals of power system economics. London: John Wiley and Sons.
2. Jigoria-Oprea, D., Lustrea, B., Borlea, I., Kilyeni, S., Andea, P., & Barbulescu, C. (2009). Short term daily load forecasting using recursive ANN. In IEEE EUROCON 2009 (pp. 631–636), St. Petersburg. doi: 10.1109/EURCON.2009.5167699
3. Quan, H., Srinivasan, D., & Khosravi, A. (2014). Short-term load and wind power forecasting using neural network-based prediction intervals, IEEE Transactions on Neural Networks and Learning Systems, 25, 303–314.
4. Selakov, A., Cvijetinovic, D., Milovic, L., Mellon, S., & Bekut, D. (2014). Hybrid PSO–SVM method for short-term load fore-casting during periods with significant temperature variations in city of Burbank, Applied Soft Computing, 16, 80–88.
5. Iran Grid Management co. (2011). http://igmc.ir. Accessed May 2016.
6. Gorunescu, F. (2011). Data mining concepts, models and techniques (Volume 12), Berlin: Springer.
7. Mosavi, A. (2014). Application of data mining in multiob-jective optimization problems, International Journal for Simulation and Multidisciplinary Design Optimization, 5, A15–319.
8. Rahman, S., & Hazim, O. (1993). A generalized knowledge-based short-term load forecasting technique, IEEE Transactions on Power Systems, 8, 508–514.
9. Senjyu, T., Mandal, P., Uezato, K., & Funabashi, T. (2005). Next day load curve forecasting using hybrid correction method, IEEE Transactions on Power Systems, 20, 102–109.
10. Paparoditis, E., & Sapatinas, T. (2013). Short-term load fore-casting: The similar shape functional time-series predictor, IEEE Transactions on Power Systems, 28, 3818–3825.
11. Abdel-Aal, R. E. (2004). Short term hourly load forecast-ing using Abductive network, IEEE Transactions on Power Systems, 19, 164–173.
12. Jain, A., & Satish, B. (2009). Clustering based short term load forecasting using support vector machines. In 2009 IEEE Bucharest PowerTech, Bucharest (pp. 1–8). doi: 10.1109/PTC.2009.5282144.
13. Espinoza, M., Joye, C., Belmans, R., & DeMoor, B. (2005). Short-term load forecasting, profile identification, and customer segmentation: A methodology based on peri-odic time series, IEEE Transactions on Power Systems, 20, 1622–1630.
14. Chicco, G., Napoli, R., & Piglione, F. (2001). Load pattern clustering for short-term load forecasting of anomalous days. In 2001 IEEE Porto Power Tech Proceedings (Cat. No.01EX502), Porto (volume 2, pp. 6). doi: 10.1109/PTC.2001.964745
15. Saini, L.M., & Soni, M.K. (2002). Artificial neural network-based peak load forecasting using conjugate gradient meth-ods, IEEE Transactions on Power Systems, 17, 907–912.
16. Fan, S., Mao, C.X., & Chen, L.N. (2005). Peak load fore-casting using the self organizing map. In Advances in neural network-ISNN 2005 (pp. 640–647), Berlin Heidel-berg: Springer.
17. Hooshmand, R., Amooshahi, H., & Parastegari, M. (2013). A hybrid intelligent algorithm based short-term load fore-casting approach, Electrical Power and Energy Systems, 45, 313–324.
18. Debnath, A., Majumder, M., & Pal, M. (2015). A cognitive approach in selection of source for water treatment plant based on climatic impact, Water Resources Management, 29, 1907–1919.
19. Bhowmik, K.L., Debnath, A., Nath, R.K., Das, S., Chattopadhyay, K.K., & Saha, B. (2016). Synthesis and characterization of mixed phase manganese ferrite and hausmannite magnetic nanoparticle as potential adsor-bent for methyl orange from aqueous media: Artificial neural network modeling, Journal of Molecular Liquids, 219, 1010–1022.
20. Debnath, A., Deb, K., Chattopadhyay, K.K., & Saha, B. (2016). Methyl orange adsorption onto simple chemical route synthesized crystalline a-Fe_2O_3 nanoparticles: Kinetic, equilibrium isotherm, and neural network model-ing, Desalination and Water Treatment, 57, 13549–13560.
21. Debnath, A., Majumder, M., Pal, M., Das, S., Chattopadhyay, K.K., & Saha, B. (2016). Enhanced adsorp-tion of hexavalent chromium onto magnetic calcium ferrite nanoparticles: Kinetic, isotherm, and neural network modeling, Journal of Dispersion Science and Technology, 37, 1806–1818.
22. Mandal, P., Senjyu, T., Urasaki, N., & Funabash, T. (2006). A neural network based several-hour-ahead electric load forecasting using similar days approach, International Journal of Electrical Power and Energy Systems, 28, 367– 373.
23. Lauret, P., Fock, E., Randrianarivony, R.N., & Manicom-Ramsamy, J.-F. (2008). Bayesian neural network approach to short time load forecasting, Energy Conversion and Management, 49, 1156–1166.
24. Fan, S., & Chen, L. (2006). Short term load forecasting based on an adaptive hybrid method, IEEE Transactions on Power Systems, 21, 392–401.
25. Torabi, M., & Hashemi, S. (2012). A data mining para-digm to forecast weather sensitive short-term energy con-sumption. In The 16th CSI International Symposium on Artificial Intelligence and Signal Processing (AISP 2012), Shiraz, Fars, pp. 579–584.
26. Escobar M. A., & Perez, L. P. (2008). Application of sup-port vector machines and ANFIS to the short-term load forecasting. In 2008 IEEE/PES Transmission and Distribu-tion Conference and Exposition: Latin America, Bogota (pp. 1–5).doi: 10.1109/TDC-LA.2008.4641839
27. Sun, C., & Gong, D. (2006). Support vector machines with PSO algorithm for short-term load forecasting. In 2006 IEEE International Conference on Networking, Sens-ing and Control, Ft. Lauderdale, FL (pp. 676–680). doi: 10.1109/ICNSC.2006.1673227
28. Niu, D., Wang, Y., & Wu, D.D. (2010). Power load forecast-ing using support vector machine and ant colony optimiza-tion, Expert Systems with Applications, 37, 2531–2539.
29. Nie, H., Liu, G., Liu, X., & Wang, Y. (2012). Hybrid of ARIMA and SVMs for short-term load forecasting, Energy Procedia, 16, 1455–1460.
30. Olson, D.L., & Delen, D.(2008). Advanced data mining techniques. Berlin: Springer.
31. Pourghasemi, H.R., Yousefi, S., Kornejady, A., & Cerda, A. (2017). Performance assessment of individual and ensemble data-mining techniques for gully erosion modeling, Science of the Total Environment, 609, 764–775.
32. Chen, W., Pourghasemi, H.R., & Naghibi, S.A. (2017). comparative study of landslide susceptibility maps

produced using support vector machine with different kernel functions and entropy data mining models in China, Bulletin of Engineering Geology and the Environ-ment, 1–18.
33. Wu, X., Kumar, V., Ross Quinlan, J., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G.J., Ng, A., Liu, B., Yu, P.S., Zhou, Z.-H., Steinbach, M., Hand, D.J., & Steinberg, D. (2008). Top 10 algorithms in data mining, Springer, Knowledge and Information Systems, 14, 1–37.
34. Acharya, N., Acharya, S., Panda, S., & Nanda, P. (2016). An artificial neural network model for a diesel engine fuelled with mahua biodiesel. In J. Kacprzyk (Ed.), Advances in intelligent systems and computing (Volume 556, pp.193–201). Poland: Springer.
35. Gupta, S., Kumar, R., Lu, K., Moseley, B., & Vassilvitskii, S. (2017). Local search methods for kMeans with outliers, Proceedings of the VLDB Endowment, 10, 757–768.
36. Nayak, J., Naik B., & Behera, H.S. (2016). Cluster analysis using firefly-based K-means Algorithm: A combined approach. In J. Kacprzyk (Ed.), Advances in intelligent systems and computing (Volume 556, pp. 55–64). Poland: Springer.
37. Fan, S., & Hyndman, R.J. (2012). Short-term load forecast-ing based on a semi-parametric additive model, IEEE Transactions on Power Systems, 27, 134–141.
38. Weather History for Bandarabbass, Iran. http://www.wunderground.com/history/. Accessed 2 February 2010.
39. Environment Canada (2003). Wind Chill Science and Equations, Retrieved 11 October 2006.
40. Steadman, R.G. (1979). The assessment of sultriness. Part I: A temperature-humidity index based on human physi-ology and clothing science, Journal of Applied Meteorol-ogy, 18, 861–873.
41. Rothfusz, L.P., Headquarters, N.S.R. (1990). The heat index equation (or, more than you ever wanted to know about heat index). Fort Worth, Texas: National Oceanic and Atmospheric Administration, National Weather Ser-vice, Office of Meteorology, pp. 90–23.
42. Han, J., & Kamber, M. (2006). Data mining: concepts and techniques (2nd edition), United States: Morgan Kauf-man, Elsevier.
43. Meshkani, A., & Nazemi, A. (2009). Introduction to data mining, Iran: Ferdowsi University.
44. Chakrabarti, S. (2009). Data mining: Know it all, USA: Morgan Kaufman, Elsevier.
45. Cottet, R., & Smith, M. (2003). Bayesian modeling and forecasting of intraday electricity load, Journal of the American Statistical Association, 98, 839–2849.
46. Hippert, H.S., Pedreira, C.E., & Souza, R.C. (2001). Neural net-works for short-term load forecasting: A review and evaluation, IEEE Transactions on Power Systems, 16, 44–55.
47. Witten, I. H., Frank, E. (2005). Data mining practical machine learning tools and techniques (3rd edition). USA: Morgan Kaufman, Elsevier.
48. Torabi, M., (2018). A Hybrid Machine Learning Approach for Daily Prediction of Solar Radiation, ecture Notes in Networks and Systems series. Springer
49. Mosavi, A., Bathla, Y., & Varkonyi-Koczy, A. (2017) Predicting the Future Using Web Knowledge: State of the Art Survey. In International Conference on Global Research and Education (pp. 341-349). Springer, Cham.